

Yelp Project:

**Optimal Restaurant Location
Identification using Yelp and
other data sets**



Linx Wu

Jiayi Du

Kevin Leugers

Pathik Rupwate

+ .
o

Team Members



Linxi Wu

Data Scientist &
Project lead



Jiayi Du

Data Scientist



Pathik Rupwate

Data Engineer



Kevin Leugers

Data Engineer

Executive Summary



■ The Problem

- Many restaurants don't have massive budgets for rich data sets and analysis on competitive landscape and local factors
- As a consulting firm, we have a limited hours and so are aiming to create a reusable solution that scales

■ The Solution

- Proof of concept tested for an example case in Austin, we help our client identify promising locations and impactful local factors to consider
- We create a data pipeline and model that can be easily applied to other cities in the United States. We also create analyses and a dashboard that can be easily repurposed

Research Questions:



- **Questions:**

- Which ZIP Codes are most attractive to open a restaurant of a specific type?
- Will crime in the local area impact our rating?
- For a given area of interest, what competition is in the area?
- What specific restaurant types are preferred in a city? Which areas are they present/absent?

Agenda

- Executive Summary
- Data Profile and Injection
- Data Cleaning
- Data Models
- Insights
- Recommendations

Data Profile

+

o

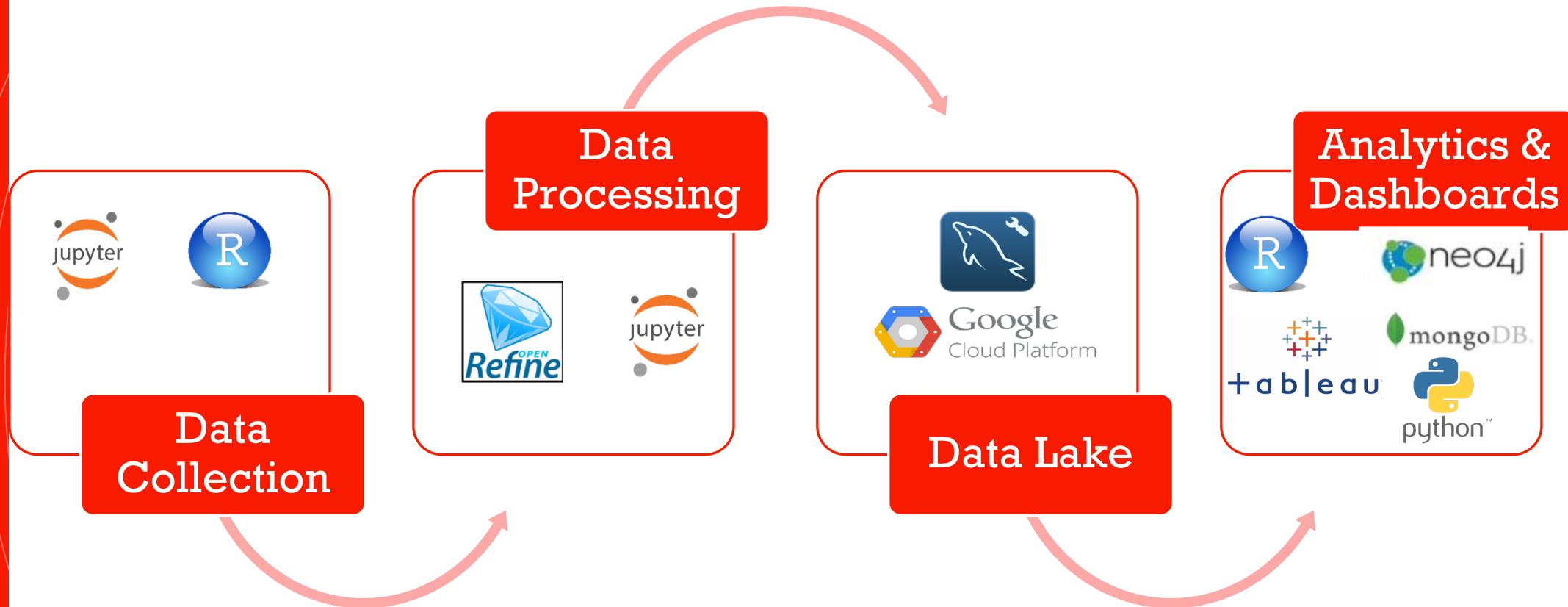
.

Data Profile



Data Source	Description	Size	Rows/Col
Austin Annual Crime Data	<i>Crime type by ZIP Code and date</i>	2.2MB	37462/5
Austin Neighborhoods by zip code	<i>Zip Codes and their associated neighborhoods in Austin</i>	5 KB	24/2
US Income	<i>Average income by ZIP Code and breakdown of population by income bracket</i>	475 KB	9738/5
Yelp business data	<i>Reviews, name, star rating, business category, lat/long, ZIP Code by business ID</i>	2.6GB	160,500/30
Unique ZIP Code List	<i>Exhaustive list of US Postal Codes</i>	350KB	33,120/1

Data Implementation Tools



Data Ingestion

- We worked with very large data sets with over 160,000 rows. The large data files were cleaned with Openrefine due to its speed and ability to export files as csv, json, sql, and text.
- We also used sql and R to transform the tables into more understandable and usable tables.
- We cleaned empty cells, renamed or dropped columns, corrected spelling errors and narrowed down on Austin, Texas as our city of choice to analyze due the large sample size of Yelp data from Austin, Texas.

In [14]:

```
!pip3 install beautifulsoup4

import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
%matplotlib inline
import re
import time
from datetime import datetime
import matplotlib.dates as mdates
import matplotlib.ticker as ticker
from urllib.request import urlopen
from bs4 import BeautifulSoup
import requests
```

OpenRefine yelp_academic_dataset_business.json [Permalink](#)

Facet / Filter Undo / Redo 0 / 0 Open... Export... Help

Reset All Remove All

Blank values per column change

51 choices Sort by: name count

- attributes 160585
- hours 160585
- attributes - RestaurantsCounterService 160545
- attributes - Open24Hours 160543
- attributes - DietaryRestrictions 160517
- attributes - AgesAllowed 160488
- attributes - HairSpecializesIn 159392
- attributes - BYOB 157066
- attributes - BYOCorkage 156918
- attributes - Corkage 156612
- attributes - Smoking 156166
- attributes - GoodForDancing 155827
- attributes - CoatCheck 155397
- attributes - BestNights 155059
- attributes - DriveThru 154547

160585 rows

Show as: rows records Show: 5 10 25 50 rows « first < previous 1 - 10 next > last »

	- categories	- business_id	- postal_code	- review_count	- hours	- attributes	- hours - Thurs	- hours - Friday	- hours - Wednesday	- hours - Tuesday	- hours - Monday
81	Gastropubs, Food, Beer Gardens, Restaurants, Bars, American (Traditional), Beer Bar, Nightlife, Breweries	6tYb2HFDyw3zjuRg0shjw	80302	86							
87	Salad, Soup, Sandwiches, Delis, Restaurants, Cafes, Vegetarian	tCbtrPZA0oilYSmHG3J0w	97218	126							
97	Antiques, Fashion, Used, Vintage & Consignment, Shopping, Furniture Stores, Home & Garden	bvN78fIM8NLprQ1a1y5dRg	97214	13							
87	Beauty & Spas, Hair Salons	oaepsyc0J17qwi8cfOWg	32763	8							

Facet Text filter Edit cells Edit column Transpose Sort... View Reconcile

Split into several columns... Join columns... Add column based on this column... Add column by fetching URLs... Add columns from reconciled values... Rename this column Remove this column Move column to beginning Move column to end Move column left Move column right



Data Cleaning

OpenRefine yelp_academic_dataset_business.json Permalink

Facet / Filter Undo / Redo 0 / 0

160585 rows Show as: rows records Show: 5 10 25 50 rows

Blank values per column change

51 choices Sort by: name count

1. Oskar Blues Taproom

2. Flying Elephants at PDX

3. The Reclamory

4. Great Clips

5. CrossFit Terminus

6. Bob Likes Thai Food

7. Escott

Orlando FL 4.5 1 28.573998 -81.3892841 Dentists, Health & 135jsh9ymlttmt69Ucp7gw 7 32804 2511

businessAcceptsCreditCards

- attributes - BusinessParking

- attributes - RestaurantsPriceRange2

- attributes - BikeParking

- attributes - WiFi

- attributes - RestaurantsTakeOut

- attributes - GoodForKids

- attributes - RestaurantsDelivery

- attributes - OutdoorSeating

- attributes - ByAppointmentOnly

- attributes - RestaurantsReservations

- attributes - RestaurantsGoodForGroups

- attributes - HasTV

- attributes - Alcohol

- attributes - Ambience

hours

- attributes - RestaurantsCounterService

- attributes - Open24Hours

- attributes - DietaryRestrictions

- attributes - AgesAllowed

- attributes - HairSpecializesin

- attributes - BYOB

- attributes - BYOCorkage

- attributes - Corkage

- attributes - Smoking

- attributes - WheelchairAccessible

- attributes - NoiseLevel

- attributes - GoodForMeal

- attributes - Caters

- attributes - RestaurantsAttire

Drop columns here to remove

Extensions Wikidata

« first < previous 1 - 10 next > last »

business_id review_count postal_code address

Yb2HFdywm3zjuRg0shjw 86 80302 921 Pearl St

bd1RPZA0olIySmHG3J0w 126 97218 7000 NE Airport Way

PN78fIM8NLprQ1a1y5dRg 13 97214 4720 Hawthorne Ave

hepsvc0J17qw18ctfOWg 8 32763 2566 Enterprise Rd

E9ugAjdw0E4-8mjG13wVA 14 30316 1046 Memorial Dr SE

AUJQNT14X3KcbzacDjsMw 169 V5V 3755 Main St

```{r}

```
rest_austin <- business%>%
 filter(city == "Austin") %>%
 mutate(categories_new = strsplit(categories, split = ","))%>%
 unnest(cols = c(categories_new)) %>%
 filter(categories_new == "Restaurants")
```


```{r}



```
rest_final <- transform(rest_austin, zipcode = as.numeric(postal_code))
```

```


```

1 db.nbh.find({})

2 .projection({})

3 .sort({_id:-1})

4 .limit(100);

5

6 db.nbh.updateMany({},

7 {\$rename:{"ZIP Code":"postal_code"}});

nbh 0.044 s 27 Docs

| | _id | postal_code | Neighborhood |
|---|------------------|----------------|------------------|
| 1 | 61b3d055353e032c | 78,757 (78.8K) | Central Austin |
| 2 | 61b3d055353e032c | 78,756 (78.8K) | Central Austin |
| 3 | 61b3d055353e032c | 78,751 (78.8K) | Central Austin |
| 4 | 61b3d055353e032c | 78,703 (78.7K) | Central Austin |
| 5 | 61b3d055353e032c | 78,705 (78.7K) | Central Austin |
| 6 | 61b3d055353e032c | 78,701 (78.7K) | Downtown |
| 7 | 61b3d055353e032c | 78,722 (78.7K) | East Austin |
| 8 | 61b3d055353e032c | 78,702 (78.7K) | East Austin |
| 9 | 61b3d055353e032c | 78,724 (78.7K) | Northeast Austin |





Data Modeling & Database Design Considerations

Storage



Google Cloud Platform Data Engineering Platforms Search products and resources

SQL Databases

PRIMARY INSTANCE

All instances > finalproject

finalproject MySQL 8.0

+ CREATE DATABASE

| Name ↑ | Collation | Character set | Type | ⋮ |
|--------------------|--------------------|---------------|--------|---|
| information_schema | utf8_general_ci | utf8 | System | ⋮ |
| mysql | utf8_general_ci | utf8 | System | ⋮ |
| performance_schema | utf8mb4_0900_ai_ci | utf8mb4 | System | ⋮ |
| Restaurants | utf8mb4_0900_ai_ci | utf8mb4 | User | ⋮ |
| restaurants_star | utf8mb4_0900_ai_ci | utf8mb4 | User | ⋮ |
| sys | utf8mb4_0900_ai_ci | utf8mb4 | System | ⋮ |

Backups

Replicas

Operations

Release Notes

GCP Selection Reasons:

- **No on Prem Resources:**
 - We don't have infrastructure and resources to maintain our own servers (among other limitations)
- **Large Data Needs:**
 - Yelp hosts an enormous amount of data.
- **Expandable Storage Needs:**
 - We anticipate expansion of our tools and data as adoption of these analyses grows so have chosen GCP.
 - Google Cloud Platform will be able to dynamically respond to our increased data needs 12

Database Design Considerations



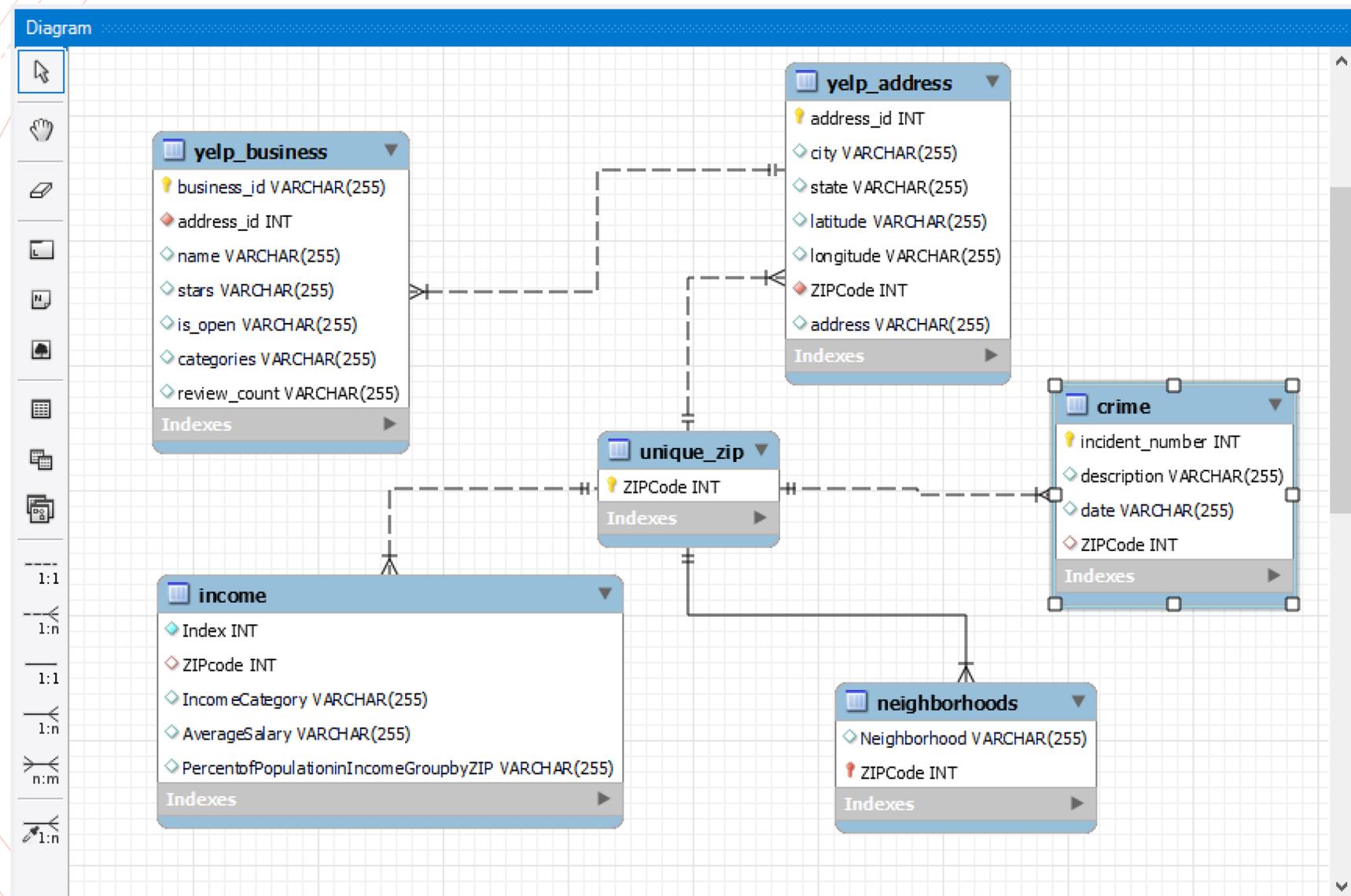
Model Selections:

- Transactional + OLAP
 - For efficient storage and future state preparation (where our client may ask to import first party data sets into our warehouse to manage), we have created a transactional model
 - However, the OLAP model will be the primary model leveraged for our reporting and analysis because the fact-dimension table set up lends itself to efficient querying

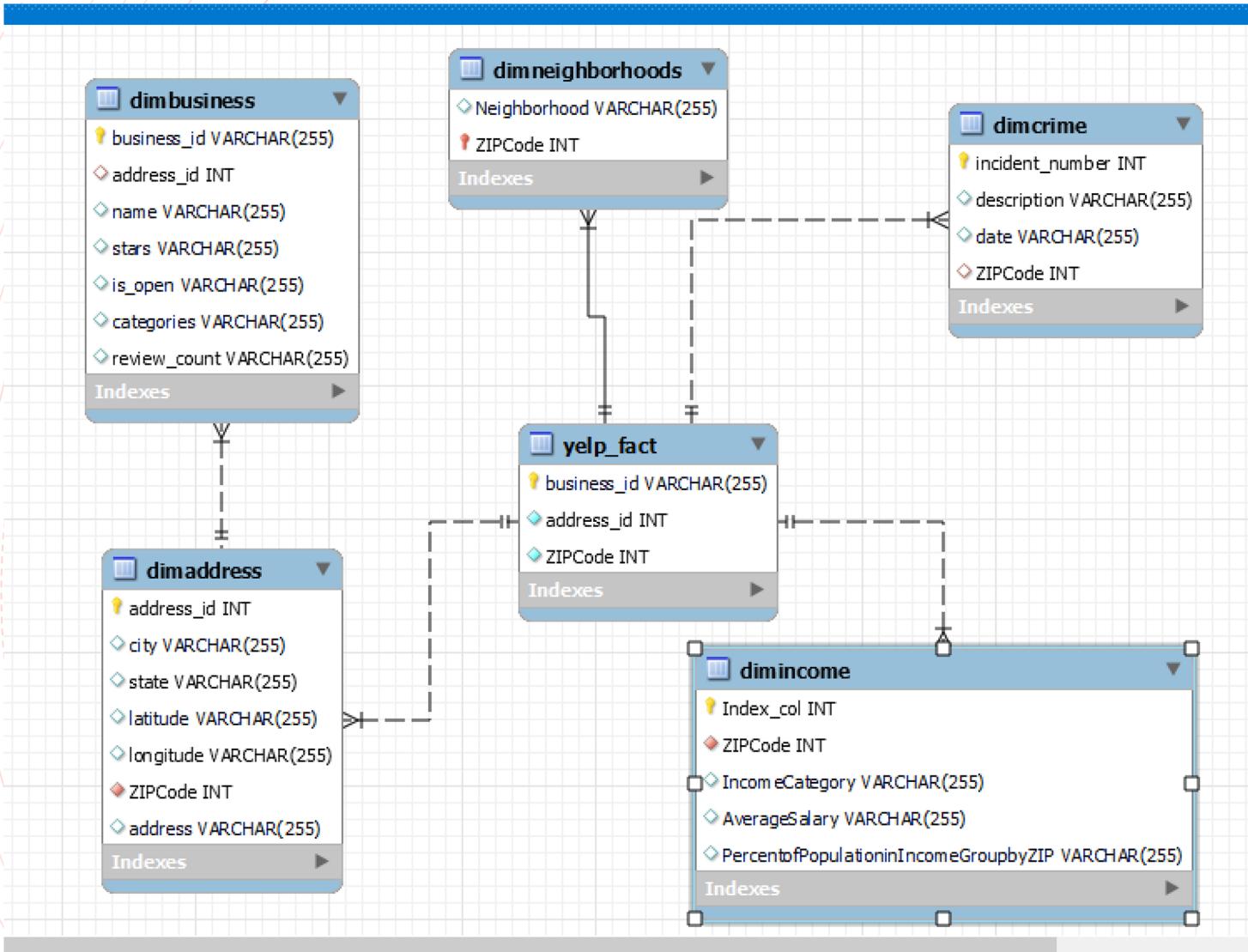
Model Creation:

- Transactional
 - 3NF: Most tables were in third normal form upon ingestion except for the Yelp data: we broke it out into a `yelp_business` table and `yelp_address` table to address dependent address fields
 - Foreign Key Connections: We brought in the unique list of US ZIP Codes to ensure the absence/presence of certain ZIP Codes in supporting tables would not exclude data or block foreign key constraints
- OLAP
 - Storage and Speed: For all data sets supporting the central Yelp data, we actively only bring in records with ZIP Codes in the Yelp data with dynamic subqueried SELECT statements during ingestion

Transactional Model



Dimensional Model



MongoDB



```
56 //Join neighbourhoods and restaurants on zip code.  
57 //We see that only 1007 out of 1303 restaurants  
58 //fall in Zip Codes in Austin Texas  
59 db.rest.aggregate([  
60   {$lookup:{  
61     from:'nbh',  
62     localField:"postal_code",  
63     foreignField:"postal_code",  
64     as: 'Neighborhood'  
65   }  
66 },  
67 {$unwind:"$Neighborhood"},  
68 {$addFields:{Neighborhood:"$Neighborhood.Neighborhood"},  
69 }]).sort({ "stars": -1 }).limit(10000);  
70
```

rest 0.332 s | 1,007 Docs

| | _id | categories | state | stars | review_count | postal_code | Neighborhood | name | longitude | latitude | is_open | city | business_id | address |
|----|----------------------|------------|-------|-------|--------------|----------------|----------------|---------------------------|-----------|----------|---------|--------|------------------------|-------------------|
| 16 | 61b3dfa8353e032c1c00 | Restaurant | TX | 5 | 7 | 78,722 (78.7K) | East Austin | DIY Thai Food | -97.7073 | 30.2965 | 0 | Austin | ImiLLRt0wlWZXs7IP3vBiA | 4209 Airport Blvd |
| 17 | 61b3dfa8353e032c1c00 | Restaurant | TX | 5 | 8 | 78,702 (78.7K) | East Austin | Mama Kong Cambodian S | -97.7164 | 30.2618 | 1 | Austin | GQVxN1JBzD2VFPxFGSx | 2211 Webber St |
| 18 | 61b3dfa8353e032c1c00 | Restaurant | TX | 5 | 12 | 78,722 (78.7K) | East Austin | Luv Fats Ice Cream | -97.7073 | 30.2965 | 1 | Austin | k6GZnr_QVIKTWsO4LNeV | 4209 Airport Blvd |
| 19 | 61b3dfa8353e032c1c00 | Restaurant | TX | 5 | 14 | 78,702 (78.7K) | East Austin | Ararat ToGo + Coffee Wind | -97.7162 | 30.2592 | 1 | Austin | iYGaz8ezrhxq5lXITMBcDQ | 2401 East 6th St |
| 20 | 61b3dfa8353e032c1c00 | Restaurant | TX | 5 | 5 | 78,701 (78.7K) | Downtown | Mike & Mike's Chicago Dog | -97.7425 | 30.2699 | 0 | Austin | LtNCzkQDZ2ybUzHdzhGqI | 722 Congress Ave |
| 21 | 61b3dfa8353e032c1c00 | Restaurant | TX | 5 | 89 | 78,757 (78.8K) | Central Austin | Hot Rod Coffee Trailer | -97.7391 | 30.3418 | 0 | Austin | BkW7DbV3_1BPItPOTWeE | 6546 Burnet Rd |

```
48 //Number of Zip codes in Austin, Texas  
49 db.rest.aggregate([  
50   {$set:{_id:"$postal_code"}},  
51   {$unionWith: {coll:"nbh", pipeline:[{$set:{_id:"$postal_code"}]}]},  
52   {$sort:{stars:1}},  
53   {$group:{_id:"$postal_code"}},  
54 ]).count();  
55
```

0.047 s

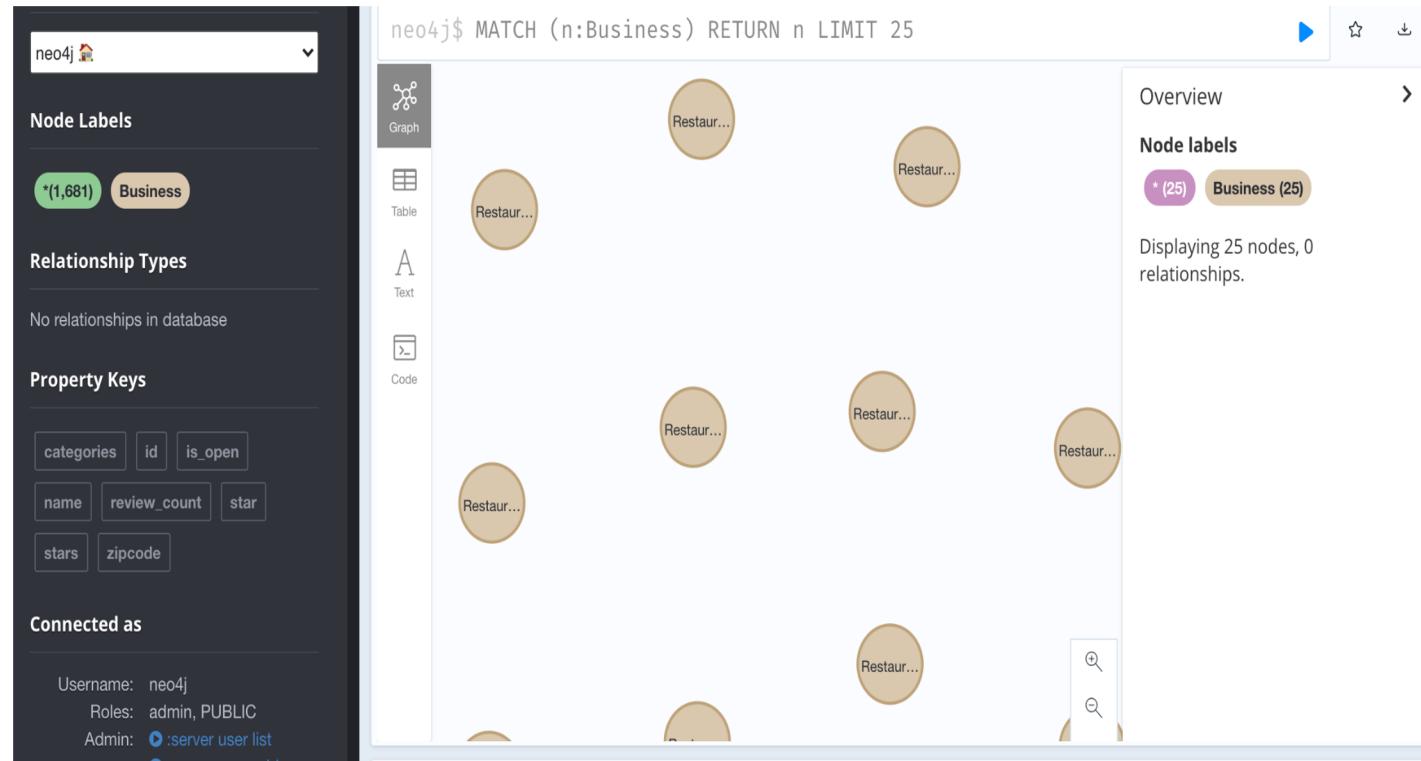
1 52

```
36 //Find all restaurants with a rating > 4 stars and sort by rating.  
37 db.rest.aggregate([  
38   {$match: {stars:{$gt:4}}},  
39   {$project:{name:1, stars:1}},  
40   {$sort : {stars:-1}}  
41 ])  
42
```

rest 0.023 s | 283 Docs

| | _id | stars | name |
|---|----------------------|-------|--------------------|
| 1 | 61b3dfa8353e032c1c00 | 5 | Little Sub Trailer |
| 2 | 61b3dfa8353e032c1c00 | 5 | Celestial Pizza |
| 3 | 61b3dfa8353e032c1c00 | 5 | SXSE Food |
| 4 | 61b3dfa8353e032c1c00 | 5 | Hawai'i Nei cafe |

Neo4j



1. Each node contain information of restaurant including star, name, zip code, review_count and is_open.
2. Each business node is in relationship with a crime node that contain crime number and a income node that contain the average income of the zip code area.

Visualization & Insights

+

o

•

Statistical Analysis

Correlation analysis

```
#Correlation analysis
```{r}
cor(correlation_data$stars, correlation_data$crime_num)
cor(correlation_data$stars, correlation_data$review_count)
cor(correlation_data$crime_num, correlation_data$salary)
```

```

```
[1] 0.1557746
[1] NA
[1] 0.1473701
```

Linear Regression

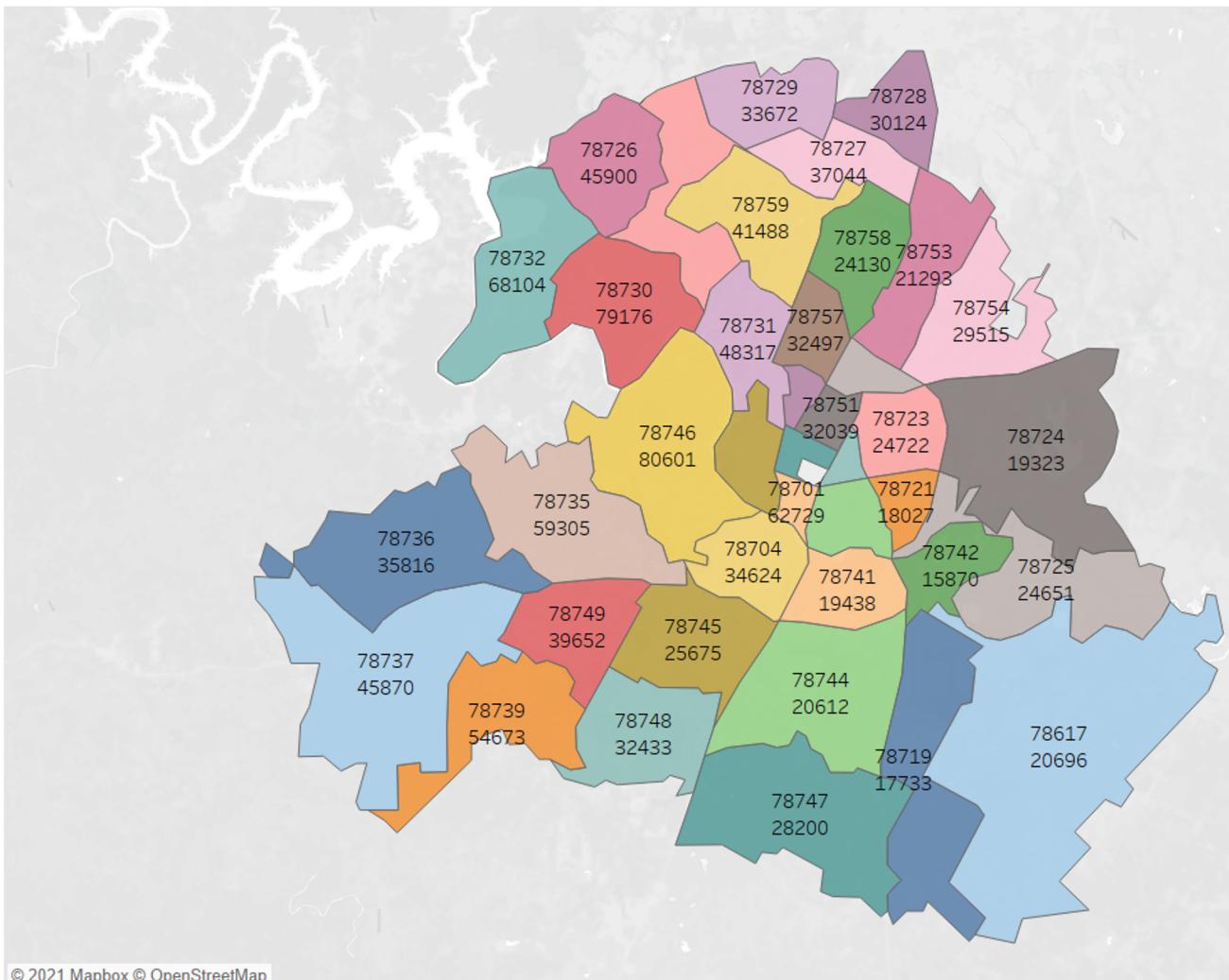
```
```{r}
model <- final %>%
 lm(stars ~ crime_num, data = .)
summary(model)
```



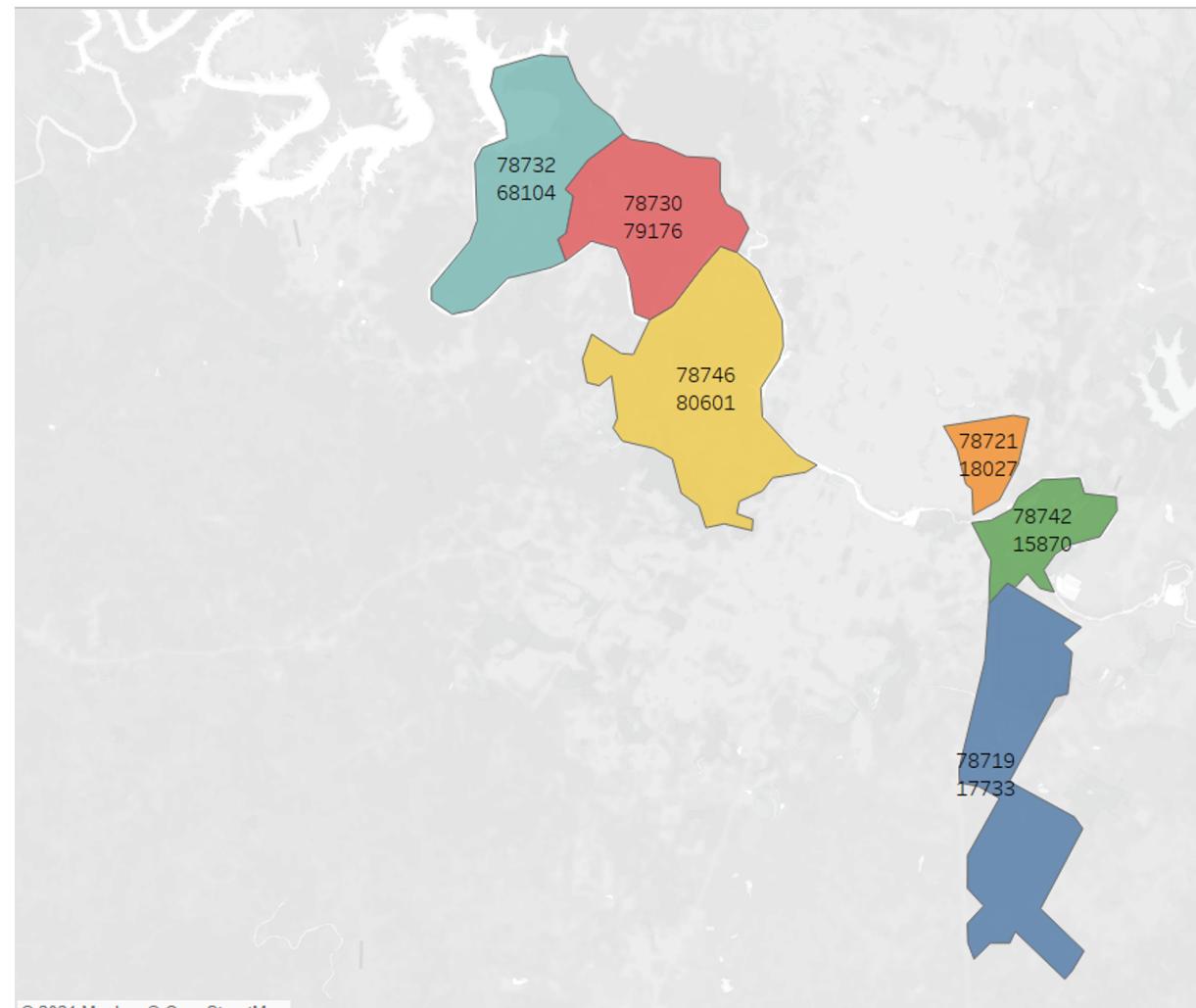
# The regions with the highest and lowest income



## Salary with Zipcode



## Salary with Zipcode



# The regions with the severe crime incidents

Crime with Number



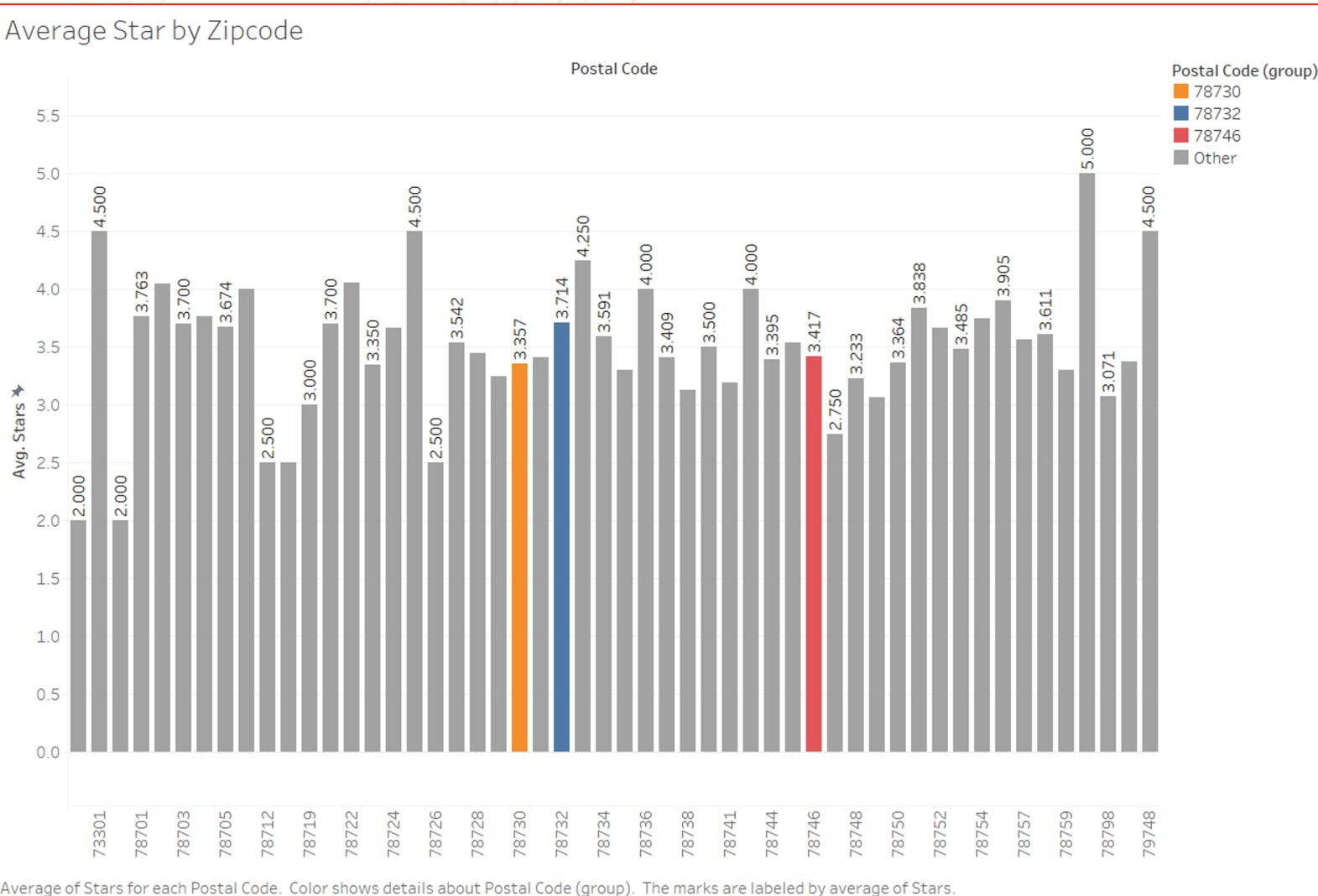
## Assumptions:

- The severe crime such as murder or assault will impact the restaurant business in the region

## Insights:

- The severe crime incidents such as murder or assault mostly happened in the downtown area
- The three regions with highest income and the three with lowest income have relatively lower severe crime rate than downtown region

# Average Star by Zip Code



## Business Logics:

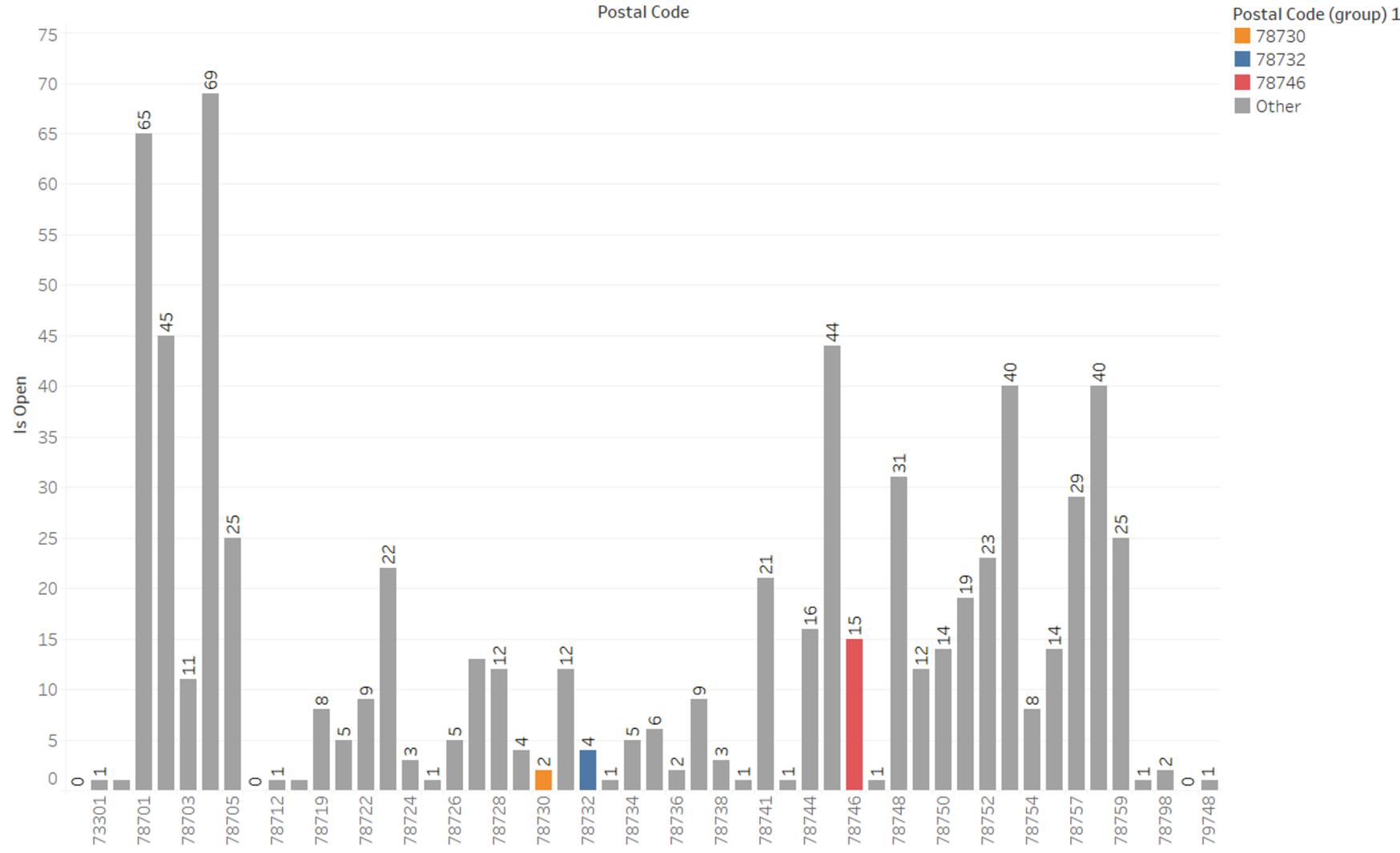
- The new restaurant opened in the regions with highest Yelp rating might suffer fierce competitions
- Finding the region with a fairly averaged rating will benefit the restaurant owners

## Insights:

- The places with the highest income level actually have a Yelp rating that fits the needs

# Number of Restaurants still opening

Number of Restaurants still opening on Yelp

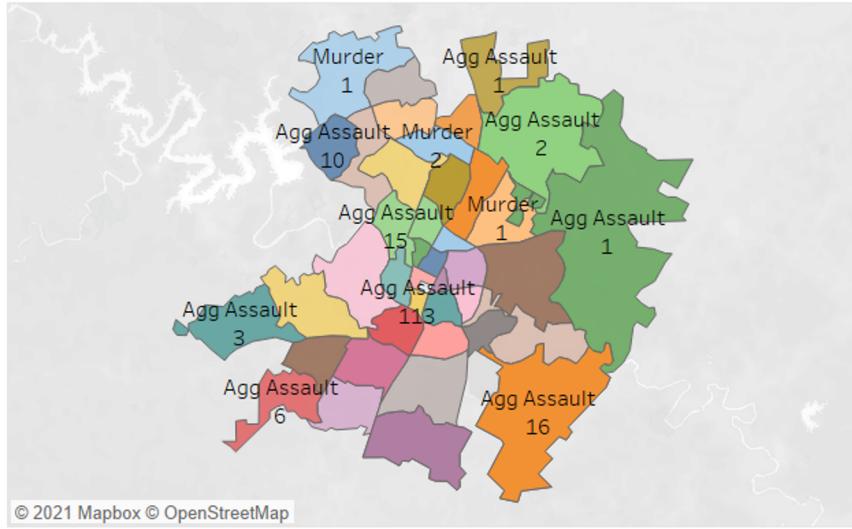


## Insights:

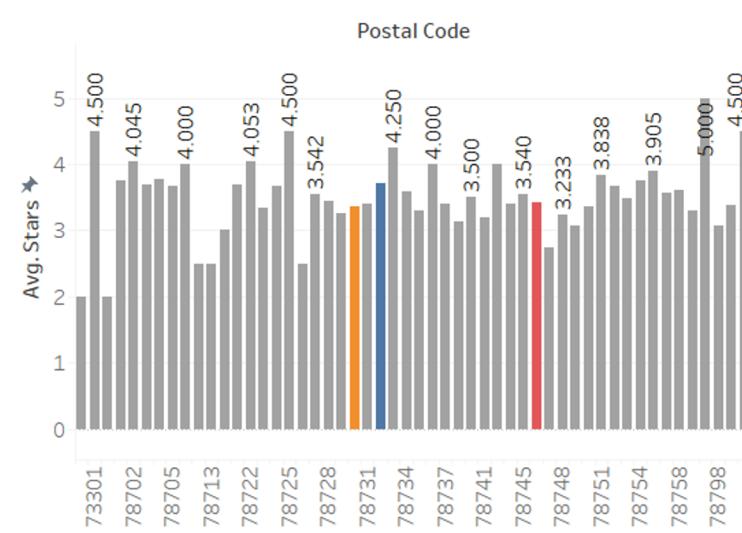
- Following the logic of last slide, we also found that the number of restaurants opening in those three regions with highest income is at the average level
- The downtown area has the highest number of restaurants opening, which indicates higher competition

# A Tableau Dashboard to help keep trends

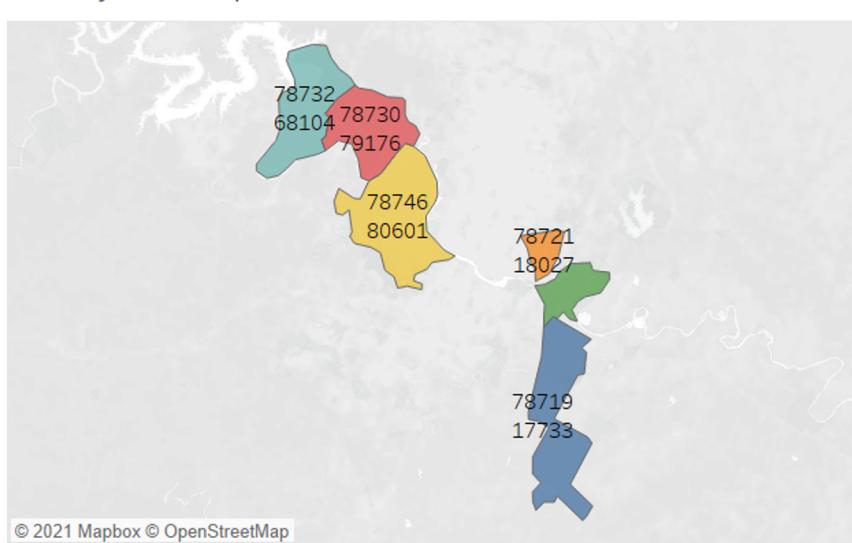
Crime with Number



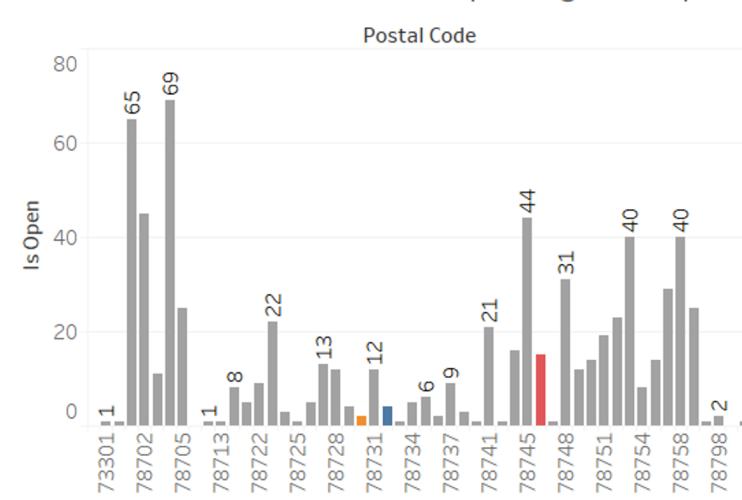
Average Star by Zipcode



Salary with Zipcode



Number of Restaurants still opening on Yelp



## Benefits:

- As the covid-19 pandemic fading away, the trends in the restaurants might change. Therefore, having a Tableau Dashboard on hand is important to make a swift change.

---

# Recommendations

+

o

•

# Recommendations

- The restaurant owners should follow the insights from the crime data, average income by zip code, and the Yelp data to make decisions
- The optimal place to open a restaurant is in region with a zip code of 78746
- The region we recommend has the highest average income, has the relatively low crime incidents, and average level of completion



A large, solid red circle is centered in the middle of the slide. Inside the circle, the words "Thank You" are written in a white, sans-serif font.

Thank You